Bruce Schneier
Harvard University

Robot Hacking Games

Hacker "Capture the Flag" has been a mainstay at hacker gatherings since the mid-1990s. It's like the outdoor game, but played on computer networks. Teams of hackers defend their own computers while attacking other teams'. It's a controlled setting for what computer hackers do in real life: finding and fixing vulnerabilities in their own systems and exploiting them in others'. It's the software vulnerability lifecycle.

These days, dozens of teams from around the world compete in weekend-long marathon events held all over the world. People train for months. Winning is a big deal. If you're into this sort of thing, it's pretty much the most fun you can possibly have on the Internet without committing multiple felonies.

In 2016, DARPA ran a similarly styled event for artificial intelligence (AI). One hundred teams entered their systems into the Cyber Grand Challenge. After completing qualifying rounds, seven finalists competed at the DEFCON hacker convention in Las Vegas. The competition occurred in a specially designed test environment filled with custom software that had never been analyzed or tested. The AIs were given 10 hours to find vulnerabilities to exploit against the other AIs in the competition and to patch themselves against exploitation. A system called Mayhem, created by a team of Carnegie-Mellon computer security researchers, won. The researchers have since commercialized the technology, which is now busily defending networks for customers like the U.S. Department of Defense.

There was a traditional human–team capture-the-flag event at DEFCON that same year. Mayhem was invited to participate. It came in last overall, but it didn't come in last in every category all of the time.

I figured it was only a matter of time. It would be the same story we've seen in so many other areas of AI: the games of chess and go, X-ray and disease diagnostics, writing fake news. AIs would improve every year because all of the core technologies are continually improving. Humans would largely stay the same because we remain humans even as our tools improve. Eventually, the AIs would routinely beat the humans. I guessed that it would take about a decade.

But now, five years later, I have no idea if that prediction is still on track. Inexplicably, DARPA never repeated the event. Research on the individual components of the software vulnerability lifecycle does continue. There's an enormous amount of work being done on automatic vulnerability finding. Going through software code line by line is exactly the sort of tedious problem at which machine learning systems excel, if they can only be taught how to recognize a vulnerability. There is also work on automatic vulnerability exploitation and lots on automatic update and patching. Still, there is something uniquely powerful about a competition that puts all of the components together and tests them against others.

Since 2017, China has held at least seven of these competitions—called Robot Hacking Games—many with multiple qualifying rounds.

To see that in action, you have to go to China. Since 2017, China has held at least seven of these competitions—called Robot Hacking Games—many with multiple qualifying rounds. The first included one team each from the United States, Russia, and Ukraine. The rest have been Chinese only: teams from Chinese universities, teams from companies like Baidu and Tencent, teams from the military. Rules seem to

Digital Object Identifier 10.1109/MSEC.2021.3121885
Date of current version: 25 January 2022

continued on p. 119

 Copublished by the IEEE Computer and Reliability Societies

 continued from p. 120

vary. Sometimes human–AI hybrid teams compete.

Details of these events are few. They're Chinese language only, which naturally limits what the West knows about them. I didn't even know they existed until Dakota Cary, a research analyst at the Center for Security and Emerging Technology and a Chinese speaker, wrote a report about them a few months ago. And they're increasingly hosted by the People's Liberation Army, which presumably controls how much detail becomes public.

Some things we can infer. In 2016, none of the Cyber Grand Challenge teams used modern machine learning techniques. Certainly most of the Robot Hacking Games entrants are using them today. And the competitions encourage collaboration as well as competition between the teams. Presumably that accelerates advances in the field.

None of this is to say that real robot hackers are poised to attack us today, but I wish I could predict with some certainty when that day will come. In 2018, I wrote about how AI could change the attack/defense balance in cybersecurity. I said that it is impossible to know which side would benefit more but predicted that the technologies would benefit the defense more, at least in the short term. I wrote: "Defense is currently in a worse position than offense precisely because of the human components. Present-day attacks pit the relative advantages of computers and humans against the relative weaknesses of computers and humans. Computers moving into what are traditionally human areas will rebalance that equation."

Unfortunately, it's the People's Liberation Army and not DARPA that will be the first to learn if I am right or wrong and how soon it matters. ■

Bruce Schneier is a security technologist, fellow, and lecturer at the Harvard Kennedy School, Cambridge, Massachusetts, USA, and the chief of security architecture of Inrupt, Inc. He blogs at www.schneier.com. Contact him at schneier@schneier.com.

Digital Object Identifier 10.1109/MSEC.2021.3136979